Modeling mitigation and adaptation policies to predict their effectiveness: The limits of randomized controlled trials

Alexandre Marcellesi and Nancy Cartwright

1. Climate policies: Mitigation and adaptation.

The negative effects of anthropogenic global warming[1] on natural and social systems promise to be diverse and important: melting of glaciers and of the polar ice caps (IPCC 2007a, 356-360) contributing to a rise of sea-levels (op. cit., 418); increase in the frequency and intensity of extreme weather events like droughts, heat waves, or floods (IPCC 2012); decrease in crop productivity resulting in increased risk of hunger (IPCC 2007b, 298); increased risk of extinction for a great number of plant and animal species (op. cit., 792); etc. Most of these negative effects are expected to occur regardless of the way emissions of greenhouse gases (GHGs) evolve in the future, and some of them are already being observed.

It is not, however, too late for policy makers to act. First, though many of the effects of global warming will inevitably occur, their intensity depends on how large the rise in average temperature turns out to be. Reducing emissions of GHGs, the cause of anthropogenic global warming, can thus help moderate the intensity of these effects. Second, because most of the

[1] We use the expressions 'anthropogenic global warming' and 'climate change' interchangeably in this paper.

effects of global warming will inevitably occur, policies for adapting to these effects and limiting their harmful consequences are necessary.[2]

This paper is about some of the serious problems we can expect to face in modeling the effects of climate change policies---in evaluating the effectiveness of policies that have been implemented and in predicting the results of polices that are proposed. The difficulties we will discuss are shared with other kinds of social and economic policies, but they can be particularly problematic for climate change policies, as we will show below. Policies for addressing climate change are commonly divided into two categories, mitigation and adaptation, corresponding to the two levels at which policy makers can address climate change.[3] The Intergovernmental Panel on Climate Change (IPCC) defines a mitigation policy as "A human intervention to reduce the sources or enhance the sinks of greenhouse gases" (IPCC 2007a, 949) and an adaptation policy as an "Adjustment in natural or human systems in response to actual or expected climatic stimuli or their effects, which moderates harm or exploits beneficial opportunities." (IPCC 2007b, 869) One can put the distinction between mitigation and adaptation in causal terms by saying that while mitigation policies are designed to reduce the causes of global warming, adaptation policies are designed to moderate its harmful effects on natural and human (or social) systems.

[2] Global warming is expected to have limited positive effects, in the short run and in some regions, for instance in the domain of timber productivity (IPCC 2007b, 289). It is also the task of policy makers to design policies for taking advantages of these positive effects.

[3] This distinction is reflected in the Fourth IPCC Assessment Report. This report treats of mitigation and adaptation in two distinct parts, though it contains a chapter on the relations between them (IPCC 2007b, chapter 18).

2. Evidence-based climate policies

Agencies which fund mitigation and adaptation policies typically want 'their money's worth'; they want to fund policies 'that work', that is policies that produce the effects they are designed to produce where and when they are implemented.[4] Claims that a given policy 'works', moreover, should be based on evidence. This idea, which is at the root of the widespread evidence-based policy movement, seems natural enough: A policy should be funded, and implemented, only if there is reasonable evidence that it will produce the desired effect in the specific location and at the specific time at which it is implemented.

In order to produce such evidence, organizations implementing policies are invited to conduct 'impact evaluations'. Impact evaluations (IEs) are studies measuring the effects of policy interventions. They are, by definition, retrospective: A policy must have been implemented for its effects to be measured. These IEs have two main functions: First, when an IE establishes that the policy had the effect it was designed to have, it thereby provides a post hoc justification for the decision to fund and implement the policy. Second, the results of IEs are supposed to inform subsequent policy decisions by providing evidence supporting predictions about the effectiveness of policies.

[4] They also want policies that have large benefit/cost ratios. We leave aside issues related to cost-benefit analysis itself in what follows, and focus on the preliminary step to any such analysis: the evaluation of the likelihood that a policy will yield the intended benefit.

Both functions are important, and this is why many of the agencies that fund policies devote part of their resources to IEs. An example in the domain of climate policies is the Global Environment Facility (GEF). The GEF, an intergovernmental agency which funds many mitigation and adaptation policies, has its own evaluation office, which produces guidelines for conducting IEs.[5]

As we mentioned above, the aim of IEs is to measure the effects of policy interventions. This is essentially an issue of causal inference. Teams of researchers that carry out IEs are, in the words of statistician Paul Holland, in the business of "measuring the effects of causes." (Holland 1986, 945) The extensive literature on causal inference in statistics and related disciplines (e.g. econometrics or epidemiology) provides policy makers with many different methods, experimental and observational, for conducting IEs.

Indeed, the counterfactual approach to causal inference (Rubin 1974, Holland 1986) which is prominent in statistics has had a palpable influence on the field of evaluation. According to the World Bank's guide to impact evaluation, for instance,

[5] See http://www.thegef.org/gef/eo_office. Other funding agencies such the World Bank (http://ieg.worldbankgroup.org/), the International Monetary Fund (http://www.ieo-imf.org), or the US Food and Drug Administration (http://www.fao.org/evaluation/) also have their own evaluation offices. There are also organizations, such as the International Initiative for Impact Evaluation (3ie, http://www.3ieimpact.org/), whose sole role is to fund and carry out IEs. The multiplication of evaluation offices results in the multiplication of guidelines and methodologies for conducting IEs.

> To be able to estimate the causal effect or impact of a program on outcomes, any method chosen must estimate the so-called *counterfactual*, that is, what the outcome would have been for program participants if they had not participated in the program. (World Bank 2011, 8, emphasis added)[6]

As this quotation hints, the idea at the root of the counterfactual approach is that the size of the contribution of a putative cause C to an effect E among program participants is identical to the difference between the value of E for those participants in a situation in which C is present and the value which *E would* take in a situation in which C is absent, all else being equal. If this difference is equal to zero, then C is not a cause of E in that population; if it is greater than zero, then C is a positive cause of E, and if it is smaller than zero, then C is a negative cause of E. According to the counterfactual approach to causal inference, answering the question 'What is the effect of C on E in a given population?' thus requires answering the following counterfactual queries 'What value would E take for individuals in that population exposed to C were C absent,

[6] It is widely assumed, and not just by the World Bank, that answering a causal question about the effect of a policy just is to answer some counterfactual question about what would have happened in the absence of the policy. Thus Duflo and Kremer, both members of the influential Jameel Poverty Action Lab at MIT, claim that, "Any impact evaluation attempts to answer an essentially counterfactual question: how would individuals who participated in the program have fared in the absence of the program?" (Duflo and Kremer 2003, 3) And Prowse and Snilstveit, in a review of IEs of climate policies, claim that, "IE is structured to answer the [counterfactual] question: how would participants' welfare have altered if the intervention had not taken place?" (Prowse and Snilstveit 2010, 233)

all else being equal?' and ‘What value would E take for individuals not exposed to C were C present, all else being equal?’

This commitment to a counterfactual approach goes together with a strong preference for experimental methods, and for randomized controlled trials (RCTs) in particular, over observational methods. According to their advocates,[7] RCTs yield the most trustworthy or, as development economists Esther Duflo and Michael Kremer put it (Duflo and Kremer 2003), "credible" estimates of the mean effect of C on E in a given population. RCTs are, to use a common expression, the 'gold standard' of causal inference.[8]

3. What are RCTs, and why are they considered the 'gold standard'?

RCTs are experiments in which individuals in a sample drawn from the population of interest are randomly assigned either to be exposed or not exposed to the cause C, where an individual can be anything from a single student to a single village to a hospital to a single country or region. Individuals who are exposed to C form the 'treatment' group while individuals who are not exposed form the 'control' group.[9] Random assignment does, in ideal circumstances and along with a sufficiently large sample, make it probable that the treatment and control groups are homogeneous with respect to causes of E besides C. And the homogeneity of the two groups with respect to causes of E other than C enables one to answer the counterfactual question 'What would be the mean value of E for individuals (in the study population) exposed to C were C

[7] Who are sometimes called 'randomistas' as in, e.g., (Ravallion 2009).

[8] See, e.g., (Rubin 2008).

[9] The terminology comes from clinical trials.

absent, all else being equal?' by citing the mean value taken by E for individuals not actually exposed to C.[10] In other words, ideally conducted RCTs make it likely, by their very design,[11] that all else is indeed equal between the treatment and control groups, and thus that the actual mean value of E for the control group can be identified with the mean value which E would take for the treatment group were individuals in this group not exposed to C (and vice-versa for the control group). This is in turn enables one to estimate the mean of the difference between the effect an individual would have were they subject to C versus were they not---often called the *causal* or *treatment effect* of C on E---in the sample, or study population, accurately.[12]

Here is a different way to put it. Assume that the effect of interest E is represented by a continuous variable Y_i and that the putative cause C is represented by a binary variable X_i taking value 1 when individual *i* is exposed to the cause and 0 when it is not. Assume also that the

[10] It also enables one to answer the question 'What would be the mean value of E for individuals (in the study population) not exposed to C were C present, all else being equal?' by citing the mean value taken by E for individuals actually exposed to C. Note that we are here talking about mean values of E over the treatment and control groups respectively. RCTs enable one to estimate the mean causal effect of C on E in a given population, not the individual causal effect of C on E for any specific individual in this population.

[11] RCTs are, in the words of (Cartwright Hardie 2012, §I.B.5.3), 'self-validating', i.e. their very design guarantees, in ideal circumstances, the satisfaction of the assumptions that must be satisfied in order for the causal conclusions they yield to be true.

[12] For more on RCTs and on the way they establish their conclusions see (Cartwright and Hardie 2012, §I.B.5) and (Cartwright 2010).

relationship between X_i and Y_i in the study population is governed by the following linear causal principle:

$$(CP)\ Y_i = a + b_iX_i + W_i$$

Here W_i is a continuous variable which represents factors that are relevant to the value of Y_i besides X_i. And coefficient b_i represents the effect of X_i on Y_i for i. Since b_i represents the individual-level effect of X_i on Y_i, the population-level mean effect of X_i on Y_i is by definition equal to Exp[b_i], where Exp[.] is the expectation operator.[13]

Randomly assigning individuals to the treatment and control groups in principle guarantees the probabilistic independence of X_i from both b_i and W_i, and this in turn enables one to accurately estimate Exp[b_i] from the difference between the expected value of the effect in the treatment group and its expected value in the control group.[14] This difference is equal to:

$$Exp[Y_i|X_i = 1] - Exp[Y_i|X_i = 0] = (a + Exp[b_i|X_i = 1] + Exp[W_i|X_i = 1])$$
$$- (a + Exp[b_i|X_i = 0] + Exp[W_i|X_i = 0])$$

In the ideal case in which assignment of individuals to either treatment or control genuinely is independent of b_i and W_i, this difference is the mean treatment effect---often referred to as just the 'treatment effect'---and can be estimated from the observed outcome frequencies. It is equal to:

[13] We treat 'mean', 'expectation' and 'expected value' as synonyms here.

[14] The probabilistic independence of X_i from b_i guarantees that the size of the effect of C on E for i is causally unrelated to whether i is assigned to the treatment or the control group. And the probabilistic independence of X_i from W_i guarantees that whether i is assigned to the treatment or control group is causally unrelated to the causes of E that do not appear in (CP).

$$Exp[Y_i|X_i = 1] - Exp[Y_i|X_i = 0] = Exp[b_i].$$[15]

So the mean treatment effect is non-zero just in case Exp[b_i] is non-zero, which can happen only if b_i is non-zero for some i in the population, which means that for that individual X_i does contribute to the value of Y_i: X_i causes Y_i in that *i*.

Experimental and observational studies in which assignment to the treatment and control groups is non-random are widely considered less desirable than RCTs because their designs, unlike that of RCTs, do not in principle make the causal homogeneity of the two groups (regarding causes of E other than C) probable, even in large samples, or, alternatively, their designs do not guarantee the probabilistic independence of X_i from b_i and W_i. This is why RCTs are considered the 'gold standard' by a large number of social and policy scientists.

If RCTs are the 'gold standard' for measuring the effects of causes, and if the aim of IEs is to measure the effects of policy interventions, then it seems legitimate to conclude that IEs should be designed as RCTs whenever possible. Indeed, this is the view advocated by a variety of policy scientists, for instance members of the Jameel Poverty Action Lab (J-PAL) such as Esther Duflo. J-PAL funds and carries out IEs that use RCTs, at the exclusion of any other evaluation methodology.[16] The view that RCTs provide the best evidence regarding the effects of policies is

[15] For the full proof see e.g. (Holland and Rubin 1988, 209-210). Essentially the same results as these hold for more complicated functional forms for (CP); we choose the linear form for ease of illustration.

[16] Though this does not mean that J-PAL members only work on RCTs, it does mean that all the IEs sponsored and conducted by J-PAL take the form of RCTs.

also embraced by the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group, a group of health scientists that produces standards for rating the quality of evidence. According to GRADE's evidence-ranking scheme, adopted by many agencies worldwide including the World Health Organization, results from RCTs are rated as having 'high quality' while results from observational studies receive a 'low quality' rating (Balshem, Helfand, Schünemann et al. 2011, 404, table 3). The views of these organizations about RCTs are echoed in hundreds of other agencies dedicated to vetting policy evaluations around the Anglophone world in areas from education to crime to aging to climate change.

So are RCTs a "silver bullet" for policy evaluation, to use an expression from (Jones 2009)? How relevant to policy making is the evidence they generate? Should the evidence base for mitigation and adaptation policies be improved by conducting RCT-based IEs? We will argue below that RCTs have important limitations and that the emphasis put on them contributes to obscuring questions that must be answered for the effectiveness of policy interventions to be reliably predicted. In §4 and §5 we will show, first in theory and then in practice---using a particular family of mitigation policies as a concrete example, that even if we agree that an RCT is necessary, results from RCTs provide only a small part of the evidence needed to support effectiveness predictions. Then, in §6, we will show that RCTs are ill-suited to evaluate the effects of most adaptation policies. Our main aim is to underline some particular methodological problems that face the use of RCTs to evaluate mitigation and adaptation policies. We use particular policy examples to illustrate these problems. But we do not aim to offer an exhaustive treatment of these particular policies nor of the full range of challenges that arise in evaluating the effectiveness of mitigation and adaptation policies in general.

4. The limited relevance of RCTs to effectiveness predictions.

4.1. Internal and external validity.

It is common, in the social and policy sciences, to distinguish between the internal and external validity of studies seeking to measure the effects of causes. According to the standard view, a study is internally valid when it produces results that are trustworthy, and externally valid when its results hold in contexts other than that of the study itself.[17] Because RCTs in principle are supposed to yield the most trustworthy estimates of treatment effects, they are also considered to have the highest degree of internal validity.[18]

It is possible for a study to have a high degree of internal validity while having a very low degree of external validity. A particular RCT, for instance, might yield conclusions that are highly trustworthy but which only hold of the study population involved in the RCT and not of any other population. Results from a study are useful for the purpose of predicting the effectiveness of policy interventions only if they are both internally and externally valid. If IEs are to be useful to policy makers, then, they must produce results that have a high degree of external validity, in addition to being internally valid.

[17] There is a lot to be said about the standard view and why the labels 'internal validity' and 'external validity' are both vague and misleading. Given limitations of space, however, these issues cannot be discussed here. For more, see (Cartwright and Hardie 2012, §I.B.6.3).

[18] The hedge 'in principle' is important. Poorly executed RCTs will not produce trustworthy estimates of treatments effects.

What does it take for a study result to be externally valid? It is often said that, for a study result to hold in contexts other than that of the study itself, the circumstances considered must be 'similar' to that of the study.[19] But what makes a set of circumstances 'similar' to some other set of circumstances? We briefly describe a framework, fully developed in (Cartwright and Hardie 2012), that enables one to address questions of external validity in a rigorous and fruitful manner.

4.2. Causal roles, causal principles, and support factors.

Causes do not produce their effects willy-nilly, at least not where it is possible to predict these effects. Rather, the effect of C on E in a given population is governed by *causal principles* that hold in that population. These causal principles can, without real loss of generality, be represented in the form of (CP) above, where C is represented by X_i and E is represented by Y_i.[20] C *plays a causal role* in (CP) just in case it genuinely appears in the equation, i.e. just in case there are values of b_i such that $b_i(X_i = 1) \neq 0$ for some i in the given population. But C does not work alone to produce a contribution to E: It works together with what we call *support factors*. These support factors are represented by b_i in (CP).[21]

---

[19] See (Cartwright and Hardie 2012, op. cit.) for a concrete example of an appeal to similarity. See also http://blogs.worldbank.org/impactevaluations/impactevaluations/why-similarity-wrong-concept-external-validity.

[20] All the conclusions we draw below apply mutatis mutandis when the relevant causal principles take more complex forms than that of (CP) (e.g. non-linear forms).

[21] You may be used to thinking of b_i as the size of the effect of X_i on Y_i. Indeed, this is the way we described it above when introducing (CP). But because, as we explain below, causes are

The idea that causes work together with support factors derives from the view that causes are INUS conditions in the sense of (Mackie 1965). To say that C is an INUS condition for E is to say that it is an *I*nsufficient but *N*ecessary part of an *U*nnecessary but *S*ufficient condition for the production of a contribution to E.[22] Mackie's classic example is that of a fire caused by a short circuit. The short circuit is not individually sufficient to produce a contribution to the fire, other factors, which we call 'support factors', are required: The presence of flammable material, the presence of oxygen, the absence of sprinklers, etc. These support factors, together with the short circuit, are jointly sufficient to produce a contribution to the fire. But they are not jointly necessary: There are other ways to contribute to a fire, i.e. there are other sets of factors---e.g. sets that have lit cigarettes instead of short circuits---that are also jointly sufficient to produce a contribution to the fire.[23]

INUS conditions, the two descriptions are equivalent: The effect of C on E just is what happens to E when C is present along with all of its required support factors.

[22] Each term in an equation like (CP) represents a contribution to the effect. Mackie's original theory does not mention 'contributions' because he only consider binary 'yes-no' variables. Our presentation is more general in that it encompasses both cases in which the cause and effect variables are binary, and more common cases in which they are not.

[23] As the 'short circuit' example makes evident, the distinction between policies and support factors is a pragmatic one. Both a policy and its support factors are causes, and so both are INUS conditions. Some factor is usually singled out as the policy because it is practical, ethically acceptable, or cost-efficient to manipulate it. Note also that we claim that all causes are INUS conditions, but not that all INUS conditions are causes.

Policies are causes, and as such are INUS conditions. They generally cannot produce a contribution to the effect they are designed to address by themselves: They need support factors. And the distribution of these support factors will differ from situation to situation. We can even expect considerable variation in which factors *are* support factors, that is which factors are needed to obtain a given effect often varies with context. Consider again Mackie's example as an illustration of this point: The short circuit may not require the absence of sprinklers in houses that are not connected to the water supply system in order to produce a contribution to the fire, though it may require the presence of a particularly large amount of flammable material in houses whose walls have been painted using fire resistant paint in order to produce the same contribution to the fire. There is no 'one size fits all' set of a support factors that, together with the cause of interest, will produce the same contribution to the effect in every context. What matters is the presence of the 'right mix' of support factors, i.e. the presence of the right support factors in the right proportions, and what the 'right mix' consists in often differs from context to context.

The framework briefly sketched above enables one to frame questions about external validity in more precise terms than does the claim that external validity is a matter of how 'similar' sets of circumstances are. To ask whether a trustworthy result from a particular study regarding the mean effect of C on E will hold in a population other than the study population is to ask:

- Does C play the same causal role in the target population as in the study population?
- Are the support factors required for C to produce a contribution to E present in the right proportions in the target population?

When both questions have positive answers, C will make a positive contribution in the target population if it does so in the study population. If either has a negative answer it is still possible that C will make a positive contribution but the RCT result is irrelevant to predicting whether it will or not---it provides no warrant for such a prediction.

4.3. Which questions do RCTs answer?

An ideal RCT for the effect of C on E will give you an accurate estimate of Exp[b_i], the mean value of b_i over individuals in the study population, or treatment effect. If this estimate is larger than 0, then you know that C makes a positive contribution to E for at least some individuals in the study population. And if this estimate is smaller than 0, then you know that C makes a negative contribution to E for at least some individuals in the study population.[24]

An ideal RCT may thus get you started on your external validity inference by providing you with some trustworthy information about the causal role C plays with respect to E in at least one population, the study population. But it gets you nowhere at all towards learning what you need to know about support factors: An ideal RCT will not tell you what the support factors are (i.e. what b_i represents) nor about individual values of b_i, i.e. about the effect of C on E for particular

[24] If this estimate is equal to 0, or very close to 0, then you cannot directly draw any conclusion about the causal role played by C in the study population because you do not know whether C is ineffective or, alternatively, its positive and its negative effects balance out. We leave this case aside here.

individuals, nor for what proportion of the study population C plays a positive, or negative, role.[25]

How much further can an ideal RCT can take you on the way to a reliable external validity inference? The short answer is: Not much further. The framework introduced above makes it clear why. First, an ideal RCT will not tell you what the causal principle governing the relationship between C and E in the study population looks like.[26] Second, an ideal RCT will not tell you what the support factors required for C to produce a contribution to E in the study population are, nor how they are distributed. Third, an ideal RCT will not tell you whether C plays the same causal role in the principles governing the production of E in the target population as in the study population. Fourth, an ideal RCT will not give you information about the support factors required for C to produce a contribution to E in the target population, nor about whether the support factors needed in the target population are the same as in the study population (which, very often, is not the case). And you need these pieces of information to produce a reliable prediction about the effectiveness of a policy.

Advocates of RCTs often reply that what is needed to overcome these limitations is more RCTs, but RCTs carried out in different locations.[27] The reasoning underlying this rejoinder seems to be

[25] See (Heckman 1992) for a further critique of the limitations of RCTs when it comes to estimating parameters that are of interest for policy making.

[26] Apart from giving you a trustworthy estimate of the value of Exp[b_i].

[27] Banerjee and Duflo, for instance, make the following claim: "A single experiment does not provide a final answer on whether a program would universally 'work'. But we can conduct a

the following: If RCTs conducted in locations A, B, and C all yield conclusive results regarding the effects of a policy, then you have strong evidence that this policy will produce the same effects when you implement it in a fourth location, call it D. This reasoning, however, is problematic insofar as it assumes without justification that the policy can play the same causal role in D as it does in A, B, or C. Since the RCTs in A, B, and C cannot individually tell you what causal principle is at work in each of these locations, their conjunction cannot, a fortiori, tell you what causal principle is at work in D. And if you don't know what causal principle is at work in D, then you also don't know whether the policy can play there the causal role you want it to play.[28]

Inferring from results in three---or even a dozen or two dozen---different locations, no matter how different they are, to the next one is a notoriously bad method of inference. It is induction by

---

series of experiments, differing in […] the kind of location in which they are conducted…"(Banerjee and Duflo 2011, 14) They add that, "This allows us to […] verify the robustness of our conclusions (Does what works in Kenya also work in Madagascar?)…" (ibid.)

[28] You may think this is an uncharitable reconstruction of the argument advanced by advocates of RCTs. But the claims they sometimes make, e.g. Banerjee and Duflo's claim, quoted in note 28, regarding the need for several RCTs in order to establish that a policy works "universally", seem to invite reconstructions that are far less charitable. One could thus see advocates of RCTs as advancing an argument of the form 'If RCTs produce conclusive results in A, B, and C, then the policy works "universally", and it will therefore work in D'. This construal seems less charitable in that it attributes to advocate of RCTs a claim (the conditional in the previous sentence) that's highly likely to be false.

simple enumeration. Swan 1 is white, swan 2 is white, swan 3 is white…. So the next swan will be white. Of course science does make credible inductions all the time. But their credibility depends on having good reason to think that the individuals considered are the same in the relevant way, that is in the underlying respects responsible for the predicted feature. In the case of causal inference from RCT populations that means that they are the same with respect to the causal role C plays and with respect to having the right mix of the right support factors.

Policy scientists writing about mitigation and adaptation policies often lament the current state of the evidence base and, naturally, call for its "strengthening" via rigorous IEs (Prowse and Snilstveit 2010, 228). So should agencies which fund and implement mitigation and adaptation policies carry out RCTs? Should the GEF, as a report of its Scientific and Technical Advisory Panel urges (STAP 2010), start designing its policies *as experiments*, and preferably RCTs, in order to improve the evidence base for climate change policies? The discussion above should make it clear that we think that RCTs are of limited relevance when it comes to producing evidence that's relevant for predicting the effectiveness of policies. We illustrate this point in the next section by examining a particular family of mitigation policies.

5. Predicting the effectiveness of mitigation policies

5.1. Mitigation via Payments for Environmental Services.

Payment for Environmental Services (PES) programs are policies that seek to conserve the environment by paying landowners to change the way they use their land. Environmental, or ecosystem, services (ESs) are loosely defined as "the benefits people obtain from ecosystems."

(MEA 2005, 26) PES policies involve a buyer, the user of the ES or a third-party acting on her behalf, and a seller, the provider of the ES.[29]

Thus a person who owns a forest and uses it for a timber activity may provide ESs by stopping this activity and by replanting trees that were cut down. In this case, the ESs provided consist in the protection of currently existing carbon stocks, via avoided deforestation, and the improvement of carbon sequestration, via the planting of new trees. Both of these ESs are directly relevant to climate change mitigation, though not all PES programs target ESs that are relevant to climate change mitigation. Many PES programs are designed with the conservation of biodiversity as their main aim.[30]

In order to stop her timber activity, the landowner described above must have an incentive to do so. Why stop her timber activity if this means a loss of earnings, and why replant trees if this means a cost without a benefit? This is where PES programs come in: They are supposed to create the incentives necessary for landowners to change the way they use their land and provide an ES. As Engel et al. put it: "The goal of PES programs is to make privatively unprofitable but

[29] In the case of mitigation-relevant PES program, the buyer of the ES often is an intergovernmental agency, e.g. the GEF, acting as a third-party on behalf of users of the ES. When the GEF is the buyer of the ES, the users it represents are the citizens of states that are members of the UN.

[30] Of course, many PES programs that target biodiversity also results in the protection of carbon stocks and, conversely, many PES programs that target climate change mitigation also result in the conservation of biodiversity.

socially-desirable practices become profitable to individual land users, thus leading them to adopt them." (Engel, Pagiola, and Wunder 2008, 670)[31]

Governmental and intergovernmental agencies see PES programs targeting deforestation as offering a major opportunity for mitigating climate change. A significant portion of the total emissions of GHGs, and CO_2 in particular, comes from deforestation.[32] If PES programs can create incentives to reduce deforestation, especially in developing tropical countries in which deforestation is a major concern, then they can contribute to a reduction in emissions of GHGs, and thus to a moderation of global warming and of its negative effects.[33]

PES programs are modeled after existing conditional cash transfer programs in domains such as development, for instance the Mexican *Oportunidades* program.[34] There are numerous IEs,

[31] The theory behind PES programs comes from the work of Ronald Coase on social cost (Coase 1960). But see (Muradian et al. 2010) for an alternative theoretical framework within which to understand PES programs.

[32] 20% according to (IPCC 2007a), 12% according to (van der Werf, Morton, DeFries et al. 2009).

[33] The UN, for instance, is developing a program called 'REDD+' that relies on PES-type programs in order to reduce deforestation. Note that 'REDD' is an acronym for 'Reduction of (carbon) Emissions from Deforestation and forest Degradation'.

[34] In the *Oportunidades* (originally PROGRESA) program, parents receive conditional payments for activities that improve human capital, e.g., enrolling their children to school. The idea is to reduce poverty both in the short-term, via the cash payments, and the in the long-run, by

including ones that take the form of RCTs, measuring the effects of conditional cash transfer programs that target poverty-reduction and education. This is particularly true for the *Oportunidades* program, first implemented in 1997 (See, e.g., Parker and Teruel 2005). This is not the case for PES programs and, in particular, for those PES programs that are relevant to climate change mitigation. There are few IEs measuring the effects of PES programs on, e.g., deforestation. And there are no completed IEs of PES programs that takes the form of an RCT.

The current state of the evidence base for PES programs is deplored by Pattanayak et al., who "see an urgent need for quantitative causal analyses of PES effectiveness." (Pattanayak, Wunder, and Ferraro 2010, 267) "Such analyses", they add, "would deliver the hard numbers needed to give policy makers greater confidence in scaling up PES." (ibid.) In this spirit, the report to the GEF mentioned above (STAP 2011) urges the intergovernmental organization to design its policies---including PES programs---as experiments as much as is possible, and this in order to facilitate the evaluation of their effects.

5.2. What will RCTs add to the evidence base for PES programs?

Responding to the call for an improvement of the evidence base for the effectiveness of PES programs in securing environmental services, MIT's J-PAL, in collaboration with the International Initiative for Impact Evaluation (3ie) and Innovations for Poverty Action (IPA), is currently carrying out an RCT aimed at measuring the effectiveness of a PES program in

---

improving human capital. The payments in this program, as well as in PES programs, are conditional in that they are made only if the service (e.g. an ES) is actually provided: They are not one-time payments that are made upfront.

reducing deforestation and biodiversity loss in the Hoima and Kibaale districts of Western Uganda.[35] Deforestation rates are particularly high in these two districts, where landowners "often cut trees to clear land for growing cash crops such as tobacco and rice or to sell the trees as timber or for charcoal production." (Jayachandran 2013a)

The design of J-PAL's RCT is as follows (Jayachandran 2013b, 311). First, 1,245 private forest owners---spread over 136 villages---were identified. They form the RCT's study population. A survey was then conducted to record several of their characteristics: number of hectares of land owned, past tree-cutting behavior, attitude toward the environment, access to credit, etc. 65 out of the 136 villages---representing 610 landowners---were then randomly assigned to the treatment group, the remaining villages being assigned to the control group. Landowners residing in villages in the treatment group were called into meetings by a local non-governmental organization (NGO), the Chimpanzee Sanctuary & Wildlife Conservation Trust (CSWCT), to receive information about the program as well as contract forms. The 'treatment' that is randomly assigned in this RCT can thus be described as 'Being offered the opportunity to sign a PES contract with CSWCT'. One of the aims pursued by J-PAL's scientists here is to estimate the effect of this treatment on deforestation and biodiversity loss.

Landowners who chose to participate in the program (or take up the 'treatment') then signed contracts with the local NGO. As Jayachandran (2013b, 311) reports,

> The contract specifies that the forest owner will conserve his entire existing forest, plus has the option to dedicate additional land to reforestation. Under the program, individuals

[35] The project is supposed to last for four years, from April 2010 through April 2014.

may not cut down medium-sized trees and may only cut selected mature trees, determined by the number of mature trees per species in a given forest patch. Participants are allowed to cut small trees for home use and to gather firewood from fallen trees.

Compliance with the contract is monitored via spot checks by CSWCT staff. Landowners who comply receive $33/hectare of forest preserved annually, an amount that was selected because it is assumed to be greater than what landowners would earn from cutting down and selling trees (other than those specified by the PES contract) for timber or charcoal, or from clearing land to grow cash crops (e.g. tobacco). As we indicated above, the assumption guiding the design of this and other PES programs is that agents will modify their behavior---here, will stop cutting down trees---if they are given the right monetary incentives to do so.

This RCT, as the official project description states, is justified by the fact that "although many PES schemes have been undertaken globally, there has not been concrete proof, emanating from scientific empirical data collected from real life PES schemes, that they are effective." (GEF 2010, 6) Note, furthermore, that this study is funded by the GEF, whose administration thus seems to be sensitive to the call for RCT-based IEs of PES programs that can deliver "hard numbers" and give "concrete proof" based on "scientific empirical data" of the effectiveness of "real life" PES programs.

As the project description indicates, one of the aims of the study is to generate, develop and disseminate a "replicable PES model based on lessons learned and best practices." (GEF 2010, 3) The aim of this RCT thus is not simply to demonstrate the effectiveness of the specific PES programs implemented in the Hoima and Kibaale districts in producing ESs. The explicit aim is

to show that PES programs aimed at reducing deforestation and biodiversity loss are effective *in general*, and to develop a PES model that can be scaled up and applied in locations besides select districts in Western Uganda.

Is the RCT currently carried out by J-PAL likely to achieve the result sought? Is it likely to provide strong evidence that PES programs work in general? How much evidence can it provide for this conclusion? If you are a policy maker contemplating the implementation of a PES program, is the RCT likely to provide reasonably strong evidence that such a program will work in the location you are targeting? We do not believe so, for reasons that were advanced in their theoretical form in §4.3. The J-PAL RCT, if it is carried out according to the script, will deliver an accurate estimate of the mean effect of the PES program on deforestation and biodiversity loss in the study population.

But it will not reveal the causal principle governing the relationship between the PES program and the reduction of deforestation and biodiversity loss in the study population.[36] It also won't tell you what support factors are needed for the PES program to play a positive causal role in the study population, nor how these factors are distributed in this population. The J-PAL RCT will not, a fortiori, tell you where the causal principle at work in the study population also holds in the population you are targeting. And it won't tell you what the support factors required for the

[36] And it won't tell you whether the same causal principle is at work in those parts of the study populations composed of landowners from the Hoima district and those parts composed of landowners the Kibaale districts.

PES program to play a positive causal role in the target population are, nor how they will be distributed.

One needs these essential additional pieces of information, regarding causal principles and support factors, in order to predict at all reliably whether the PES program will play the same causal role when it is implemented in other locations, e.g. when it is scaled up to other districts in Western Uganda, or when it is implemented in Eastern Uganda, or when it is implemented in other countries in sub-Saharan Africa, etc. One cannot arrive at a "replicable PES model", i.e. at a PES model that will work in many locations, without a detailed understanding of how the PES program works in the original study population. Nor is it clear that there is a reliable "replicable PES model" that works 'in general' to be found. It is not obvious that one can formulate substantial and useful generalizations about PES programs across settings (cultural, political, economic, religious, etc.) and, especially, across types of ESs (Can one generalize results obtained in a context in which the ES is avoided deforestation to a context in which the ES is the preservation of water resources?). The framework introduced above is designed to help you think about how a policy works when it does, and about what it would take for it to work in a different location.

We are obviously not claiming that nothing will have been learned during the four years of the J-PAL project described above, besides an estimate of some treatment effect. The policy scientists carrying out J-PAL's RCTs are neither blind nor stupid. They will gain a wealth of new knowledge regarding the local institutional and social context, the way landowners respond to the PES program, differences between villages that are relevant to the effect of the program, etc.

Note, however, that this context-specific knowledge (1) may well have been acquired even if enrollment in the PES program had not been randomly offered to landowners, (2) is just as important as is knowledge of the treatment effects to predicting the effectiveness of subsequent PES programs, and (3) is likely to be overshadowed by the "hard numbers", i.e. the estimates of treatment effects. The framework introduced above, and fully developed in (Cartwright and Hardie 2012), shows why this context-specific knowledge is essential to predicting the effectiveness of policies. And it also gives you the tools to articulate this knowledge in ways that make it relevant to effectiveness predictions.

The bottom line, here, is that if you are a policy maker contemplating the implementation of a PES program for reducing deforestation and biodiversity loss in a particular location, the results from J-PAL's RCT will offer you some guidance, but not much. You need knowledge about the causal principles at work and the support factors required for the PES program to produce a positive contribution in the location you are targeting. Let us further illustrate the importance of support factors by looking at five hypothesized support factors needed by PES programs in some locations.

5.3. Some of the support factors (sometimes) needed by PES programs.

We briefly list below five of the factors identified in the literature as playing a role in determining the effectiveness of PES programs in reducing deforestation and biodiversity loss.[37] As we noted above (§4.2), a policy might require different support factors in different contexts in

[37] See e.g. (Pattanayak, Wunder, and Ferraro 2010), (Pirard, Billé, and Sembrés 2010), (Alix-Garcia, de Janvry, Sadoulet, and Torres 2009), (GEF 2010, 35), or (Jayachandran 2013b).

order to produce the intended contribution to the effect of interest. These five factors, therefore, may be support factors for PES programs in some contexts, but not in others. The second factor---the low cost of enforcing PES programs---for instance, may not be a required support factor in contexts in which the sellers of the ES tend to abide by contracts for cultural or religious reasons.

Our framework makes it plain why these factors matter and why having evidence about their presence and distribution is crucial. If we make the unrealistic assumption that these factors are support factors always required by PES programs then, for your effectiveness prediction regarding a PES program to be properly supported by evidence, you must have evidence that these factors are present, and distributed in just the right way, in the location in which the program is to be implemented.[38] Below we list the five factors we have seen cited in the literatures about PES programs and some of the questions they immediately give rise to. But behind these there are bigger questions that need answering: 'Are these necessary in all cases?', 'What else is necessary in any particular case?', 'Will the necessary factors be in place, or can they be put in place, in the new place?', and very importantly, 'What kinds of study can help us find out the answers to these bigger questions?'

1. Strong property rights. A PES program, it is argued, can only be effective if there exists property rights and the means to enforce them in the location in which the program is to be implemented. There is no landowner for the ES buyer to sign a contract with if there is no landowner to start with. But how strong do these property rights need to be,

[38] And if the assumption that these factors are always required is dropped, then you also need evidence that these factors are indeed support factors needed for the PES program to produce the intended contribution to the effect in the location you are targeting.

and do they need to be guaranteed by a government? Where are property rights strong enough, and where are they too weak for PES programs to be effective?

2. *Low cost of monitoring and enforcing PES contracts*. If the economic and political cost of monitoring and enforcing PES contracts is high then there is an incentive for the buyer not to do so, and thus for the seller to breach the contract. These costs must be low for PES programs to be effective. But how low must they be? And how does one assess these costs?

3. Sustainable and flexible funding source. PES programs can only be effective, it is argued, if they are funded on the long-term and if the funding source is flexible enough to allow for re-negotiation of PES contracts. If the price of timber rises, then the payment for forest conservation provided to a forest owner must rise for the incentives to stay the same, and for the forest owner to keep providing an ES. Can NGOs provide sustainable and flexible funding? What about governmental agencies in countries that are politically unstable?

4. Absence of leakage. If a forest owner agrees to stop her timber activity on a parcel she owns and for which the PES contract was signed, but then goes on to use the extra earnings from the contract to buy a similarly-sized parcel nearby and resume her timber activity on that parcel, then the PES program is not effective in reducing deforestation and biodiversity loss. Opportunities for 'leakage' must be limited for the PES program to play the expected causal role. How does one assess opportunities for leakage?

5. *Access to credit*. If a forest owner cannot easily borrow money to cover emergency expenses (e.g. medical bills), then she might cut down and sell trees instead, even if she signed a PES contract covering those trees. An easy access to credit might thus lower the

chances that forest resources will be used as a ‘safety net’ and thus have a bearing on the effectiveness of the PES program. But how exactly does one measure ‘access to credit’, and how easy must access to credit be in order for the resources covered by the PES contract to stop being a ‘safety net’?

We emphasize that these are just five among the numerous factors that may be support factors required for a PES program to produce a contribution to the reduction of deforestation. The point we want to illustrate here is that J-PAL's RCT will not tell you whether these are support factors required in the location you are targeting, nor whether they are actually present there, nor how they are distributed. Unfortunately, you need this information in order to accurately predict whether a PES program will play the causal role you want it to play in the location in which you are contemplating its implementation.

6. Evaluating the effects of adaptation policies: The limits of RCTs.

Remember that adaptation policies seek to modify natural or human systems in order to reduce their vulnerability to weather-related events due to climate change. The term ‘vulnerability’ has a precise meaning in this context. According to the IPCC’s definition, the vulnerability of a system (usually some geographical unit, e.g. a city) to climate change is the “degree to which [it] is susceptible to, and unable to cope with, adverse effects of climate change, including climate variability and extremes.” (IPCC 2007b, 883) More precisely, the vulnerability of a system is “a function of the character, magnitude, and rate of climate change and variation to which [it] is exposed, its sensitivity, and its adaptive capacity.” (ibid.) An adaptation policy is designed to reduce the vulnerability of a system by reducing its sensitivity---i.e. the extent to which it is harmed by climate change---or by enhancing its adaptive capacity---i.e. its ability to adjust to

moderate the harmful effects of climate change. A distinction is often drawn between environmental vulnerability---as measured for instance by the country-level Environmental Vulnerability Index (EVI)---and social vulnerability---as measured for instance by one of the Social Vulnerability Indices (SoVi).[39]

There are various obstacles to the use of RCT-based IEs to evaluate the effects of adaptation policies. First, adaptation policies take a wide variety of forms, many of which simply do not lend themselves to randomization. Consider for instance the 'Adaptation to Climate Change through Effective Water Governance' policy under implementation in Ecuador that aims to improve the country's adaptive capacity by mainstreaming "climate change risks into water management practices..." (GEF 2007, 2) This policy will change water management practices in Ecuador, e.g. by incorporating climate risks in the country's 'National Water Plan'. How is one to evaluate the extent to which such a policy will improve Ecuador's adaptive capacity and thus reduce its vulnerability, both environmental and social, to climate change? RCTs are no help here, given that the policy is implemented at the level of an entire country. One cannot, for a variety of reasons (political, practical, etc.), randomly assign countries to particular policy regimes.

[39] See http://www.vulnerabilityindex.net/ for the EVI and http://webra.cas.sc.edu/hvri/ for the US county-level SoVI. Note two difficulties with using these indices to evaluate the effects of adaptation policies. First, they are measures of vulnerability to environmental hazards in general, whether or not they are due to climate change. Second, there is no wide consensus as to how to measure overall vulnerability (at various geographical scales), and neither is there a consensus regarding how to measure an important component of vulnerability, namely adaptive capacity.

The same point applies to the many adaptation policies that aim to improve some country's adaptive capacity, and thus reduce its vulnerability, by modifying its institutions. Here is another example. The government of Bhutan is, with the help of the United Nations Development Programme (UNDP), implementing the 'Reducing Climate Change-Induced Risks and Vulnerabilities from Glacial Lake Outburst Floods [GLOFs]' policy which, among other things, aims to integrate the risk of GLOFs due to climate change occurring in the Punakha-Wangdi and Chamkhar valleys in Bhutan's national disaster management framework.[40] Such policies, because they target country-level institutions, cannot in practice be evaluated using RCT-based IEs. The problem here is that a vast number of adaptation policies fall into this category. Note also that such policies, by their very nature, are tailored to the institutions of a particular country and so may not be implementable in any other country. A policy that improves Bhutan's adaptive capacity, for instance, may not be applicable, and a fortiori may not have the same beneficial effects, in a country which faces similar risks but has a different institutional structure (e.g. Canada, which, unlike Bhutan, is a federal state).

Second, for many adaptation policies, RCT-based IEs are superfluous. Consider for instance the Kiribati Adaptation Program (Phase II) implemented between 2006 and 2010 that included the construction of a 500 meters long seawall to protect the country's main road, a coastal road

[40] See http://www.adaptationlearning.net/bhutan-reducing-climate-change-induced-risks-and-vulnerabilities-glacial-lake-outburst-floods-punakh.

around Christmas Island.[41] One does not need an RCT in order to determine whether this seawall is helping protect the road and reduce beach erosion (inside this wall). The physical configuration of seawalls guarantees that they will reduce the sensitivity of the systems inside them to the consequences of climate change (e.g. to rising sea levels, erosion, and extreme weather events). One might argue that an RCT would enable one to determine *by how much* the Kiribati seawall reduces the sensitivity of the systems it helps protects, i.e. would enable one to estimate the size of the effect of this seawall on sensitivity. In this case, as with most adaptation policies, however, the need for an immediate reduction in sensitivity trumps the need for precise estimates of treatment effects.

One could have conducted an RCT in which the coastline along the Christmas Island road is divided into *n* sections, half of them randomly assigned to the 'seawall' group and half of them to the 'no seawall' group, and compared the condition of the road and the extent of beach erosion between sections in the 'seawall' group and those in the 'no seawall' after a year, for instance. This would have provided one with estimates of the effect of seawalls on road condition and beach erosion on Kiribati's Christmas Island (assuming both road condition and beach erosion can be reliably measured). Conducting such an RCT would make little sense for Kiribati's policy makers, however. Roads are useful only if they enable you to get somewhere, and they can only do so if they are uninterrupted and in good condition rather than irreversibly damaged at random intervals. The aim of this hypothetical example is not to caricature the position of those who, like members of the GEF's Scientific and Technical Advisory Panel (STAP 2010), call for more

[41] See http://www.thegef.org/gef/greenline/july-2012/preparation-adaptation-and-awareness-kiribati%E2%80%99s-climate-challenge.

RCT-based IEs of adaptation and mitigation policies. It is simply to illustrate that such calls sometimes conflict with the goals the policies that are to be evaluated are supposed to achieve. What matters in the end is that these policies produce the beneficial effects they were designed to produce, not that we have highly trustworthy point estimates of the size of these effects.

This is not to say that there are no adaptation policies the effects of which can be evaluated using RCT-based IEs. Policies which offer farmers rainfall index insurance, i.e. policies that insure farmers against both deficits and excesses in rainfall, can be considered adaptation policies, and their effects on the vulnerability of particular study populations to climate change can in principle be evaluated using RCTs, even though no such RCT has been conducted to date.[42] This is true in general of adaptation policies that do not seek to reduce a country's vulnerability by modifying its institutions (e.g. by incorporating climate risks into its planning tools) or its infrastructures (e.g. by building seawalls) but rather target units (e.g. individual farmers or villages) that can more easily be randomly assigned to some treatment group. The mistake here would be to think that such policies should occupy a privileged position in the portfolio of policies available to policy makers preoccupied with adapting to climate change simply because they can be evaluated using RCT-based IEs. As we showed in §5 for PES policies aiming at mitigation, the fact that a policy lends itself to randomization does not imply that it can more

[42] RCTs conducted about weather insurance usually attempt to estimate the effects of such insurance on investment decisions (see e.g. Giné and Yang 2009) or to understand the causes of weather insurance take-up (see e.g. Cole et al. 2013). See (De Nicola 2011) for a non-randomized evaluation of the effects of rainfall index insurance on the welfare of farmers and so on their adaptive capacity.

easily be generalized beyond the study population. And it also does not imply that this policy is more effective than other policies that cannot be similarly evaluated. A policy that offered Ugandan farmers the possibility of using drought-resistant seeds might lend itself to an RCT-based IE more easily than does Uganda's national irrigation masterplan,[43] but this obviously does not mean that the former is more effective than the latter at reducing the sensitivity of Ugandan farmers to droughts due to climate change.

We showed in §5 that results from RCT-based IEs of mitigation policies such as PES programs provide only a small part of the total evidence needed to support effectiveness predictions. The situation is more challenging even in the case of adaptation policies, since many of these cannot be evaluated using RCTs in the first place. The lesson of this section thus is that, both for evaluating past adaptation policies and for supporting predictions regarding the effectiveness of future adaptation policies, we need more than RCTs. Nor is it especially the issue of random assignment that raises difficulties. We face here rather problems that are endemic with comparative group studies: They are often not possible and they tell us only a little of what we need to know to make use of their own results.

6. Conclusion.

Should J-PAL scientists pack their bags and cancel the RCT they are currently carrying out in Western Uganda? No. Are RCTs a bad tool for causal inference? No. Are estimates of treatment effects irrelevant for policy making in the domain of climate change policies? No.

[43] See www.mwe.go.ug.

We want to emphasize that our criticisms are not directed at RCTs per se. Criticizing RCTs in principle makes little more sense than criticizing hammers in principle. Both RCTs and hammers are well-designed tools. One can criticize their instances: There are bad hammers and poorly conducted RCTs. And one can criticize the use to which they are put. It is the use to which RCTs are frequently put that we target and criticize.

Calling for more and more RCTs in order to strengthen the evidence base for mitigation policies such as PES programs is a bit like calling for the use of more and more hammers in order to carve a statue out of a block of marble. What one needs is not more and more hammers, but hammers and chisels, i.e. tools of a different kind. In the policy case, what one needs is not more estimates of treatment effects produced by more RCTs. If one starts with an RCT, what one needs is evidence of a different kind, evidence that is relevant to external validity inferences, and so to prediction about the effectiveness of particular policies implemented in particular contexts. The framework sketched above in §4.2 tells you what kind of evidence is needed, namely evidence about causal principles and support factors.

What we advocate corresponds, to some extent, to what Pattanayak, Wunder and Ferraro (2010, 6) call "economic archeology", i.e. the qualitative evaluation of existing policies in order to reveal the contextual factors that are relevant to their effectiveness. What we argue is that calls for an improvement of the evidence base for PES programs, and mitigation and adaptation policies in general, should emphasize the need for more "economic archeology" just as much, or even more, than they emphasize the need for estimates of treatment effects generated by RCTs.

This is particularly true for adaptation policies since, as we showed in §6, these often cannot be evaluated using RCTs. The "hard numbers" produced by RCTs---when and where they are available---are of little use for policy without knowledge of the networks of factors that give rise to these numbers, and without models of these networks (see Cartwright, forthcoming). The framework sketched here, and fully developed in (Cartwright and Hardie 2012), provides one with the means to do "economic archeology" where RCTs are involved in a rigorous and fruitful manner.

But it is important to stress that we do not need to start with RCTs in order to pursue economic archeology. The issue of course is how to do economic archeology in anything like a rigorous and reliable way. This involves understanding how best we can provide evidence about causal relations in the single case. So, besides a call for more and more RCTs, surely there should be an equally urgent call for more systematic study of what counts as evidence for causality in the single case.

Acknowledgements

Both authors would like to thank the Templeton Foundation's project 'God's Order, Man's Order and the Order of Nature', the UCSD Faculty Senate and the AHRC project 'Choices of evidence: tacit philosophical assumptions in debates on evidence-based practice in children's welfare services' for support for the research and writing of this paper. Nancy Cartwright would in addition like to thank the Grantham Research Institute on Climate Change and the Environment at LSE.

References

Alix-Garcia, J., A. de Janvry, E. Sadoulet, and J.M. Torres. 2009. Lessons learned from Mexico's payment for environmental services program. In *Payment for environmental services in agricultural landscapes*, eds. L. Lipper, T. Sakuyama, R. Stringer, and D. Zilberman, 163-188. New York: Springer.

Balshem, H., M. Helfand, H. Schünemann et al. 2011. GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology* 64: 401-406.

Banerjee, A., and E. Duflo. 2011. *Poor economics: A radical rethinking of the way to fight global poverty*. New York: PublicAffairs.

Cartwright, N. 2010. What are randomised controlled trials good for? *Philosophical Studies* 158:59-70.

Cartwright, N. Forthcoming. Will Your Policy Work? Experiments vs. Models. In *The Experimental Side of Modeling*, eds. I. F. Peschard and B. C. van Fraassen, Chicago: University of Chicago Press.

Cartwright, N., and J. Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. New York: Oxford University Press.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1-44.

Cole, S., Giné, X., Tobacman, J., et al. 2013. Barriers to Household Risk Management: Evidence from India. *American Economic Journal: Applied Economics* 5: 104-35.

De Nicola, F. 2011. The Impact of Weather Insurance on Consumption, Investment, and Welfare. Working paper.

Duflo, E., and M. Kremer. 2003. Use of Randomization in the Evaluation of Development Effectiveness. Working paper.

Engel, S., S. Pagiola, and S. Wunder. 2008. Designing payments for environmental services in theory and practice: An overview of the issues. *Ecological Economics* 65: 663-674.

GEF. 2007. Adaptation to Climate Change through Effective Water Governance in Ecuador. Project executive summary.

GEF. 2010. *Developing an experimental methodology for testing the effectiveness of payments for ecosystem services to enhance conservation in productive landscapes in Uganda*. Washington, DC: Global Environment Facility.

Giné, X., and D. Yang. 2009. Insurance, Credit, and Technology Adoption: Field experimental Evidence from Malawi. *Journal of Development Economics* 89: 1-11.

Heckman, J. 1991. Randomization and social policy evaluation. NBER technical working paper #107.

Holland, P. 1986. Statistics and causal inference (with discussion). *Journal of the American Statistical Association* 81: 945-970.

Holland, P., and D. Rubin. 1988. Causal inference in retrospective studies. *Evaluation Review* 12: 203-231.

IPCC. 2007a. *Climate change 2007: The physical science basis*. New York: Intergovernmental Panel on Climate Change.

IPCC. 2007b. *Climate change 2007: Impacts, adaptation and vulnerability*. New York: Intergovernmental Panel on Climate Change.

IPCC. 2012. *Managing the risks of extreme events and disasters to advance climate change adaptation*. New York: Intergovernmental Panel on Climate Change.

Jayachandran, S. 2013a. Evaluating a Payments for Ecosystem Services program in Uganda. Blog post from 2013/04/22 on *climate-eval* (http://www.climate-eval.org/?q=print/2235), accessed on 2013/07/17.

Jayachandran, S. 2013b. Liquidity Constraints and Deforestation: The Limitations of Payments for Ecosystem Services. *American Economic Review* 103: 309-313.

Jones, H. 2009. The 'gold standard' is not a silver bullet for evaluation. *ODI Opinion* 127.

Mackie, J. 1965. Causes and Conditions. *American Philosophical Quarterly* 2: 245-64.

MEA. 2005. *Ecosystems and human well-being: Current state and trends*. Washington, DC: Millennium Ecosystem Assessment.

Parker, S., and G. Teruel. 2005. Randomization and social program evaluation: The case of Progresa. *Annals of the American Academy of Political and Social Science* 599: 199-219.

Pattanayak, S., S. Wunder, and P. Ferraro. 2010. Show me the money: Do payments supply environmental services in developing countries? *Review of Environmental Economics and Policy* 4: 254-274.

Pirard, S., R. Billé, and T. Sembrés. 2010. Questioning the theory of Payments for Ecosystem Services (PES) in light of emerging experience and plausible developments. *Analyses* (IDDRI-Sciences Po) 4: 5-22.

Prowse, M., and B. Snilstveit. 2010. Impact evaluation and interventions to address climate change: A scoping study. *Journal of Development Effectiveness* 2: 228-262.

Ravallion, M. 2009. Should the randomistas rule? *The Economists' Voice* 6: 1-5.

Rubin, D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66: 688-701.

Rubin, D. 2008. Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association* 103: 1350-1353.

STAP. 2010. *Payments for environmental services and the global environment facility: A STAP advisory document*. Washington, DC: Scientific and Technical Advisory Panel.

van der Werf, G., D. Morton, R. DeFries, et al. 2009. CO_2 emissions from forest loss. *Nature Geoscience* 2: 737-738.

World Bank. 2011. *Impact evaluation in practice*. Washington, DC: The World Bank.